

The methodology for assessing the impact of appraisal on teachers' professional development

Aidana Shilibekova¹, Saule Vildanova¹, Venera Mussarova¹, Baurzhan Yessingeldinov²,
Moldir Ablayeva¹, Amina Kaldybek³

¹National Center for Teacher Professional Development "Orleu", Astana, Kazakhstan

²Limited Liability Partnership, Ascent Research Group, Astana, Kazakhstan

³Autonomous Educational Organization, Nazarbayev University, Astana, Kazakhstan

Article Info

Article history:

Received Jul 1, 2024

Revised Nov 10, 2024

Accepted Dec 3, 2024

Keywords:

Appraisal

Argument based approach

Assessment validity and reliability

Kane

Teacher professional development

Teacher qualification categories

ABSTRACT

In Kazakhstan, the teacher appraisal processes intended to support professional development frequently fall short of their objectives because of an excessive focus on test outcomes. This focus distorts the purpose of the appraisal, leading to a misalignment between the assessment outcomes and the actual improvement of educational practices. Addressing this critical issue, this study proposes a methodology based on Kane's argument-based approach to validity, aimed at more accurately assessing the impact of appraisal on teachers' professional development. By applying Toulmin's model of argumentation, the validity and reliability of the existing appraisal procedures were assessed, allowing key factors influencing their effectiveness to be identified. The methods also included reviewing the appraisal documents (professional standards, appraisal rules, and teacher qualification requirements) for data triangulation. The findings reveal that the proposed methodology enhances the objectivity and fairness of teacher qualification evaluations and supports meaningful professional growth. By ensuring a more comprehensive and evidence-based assessment, this methodology can improve the transparency and effectiveness of the appraisal process, ultimately contributing to higher educational standards in Kazakhstan. The study also offers practical recommendations for implementing the methodology across different levels of the education system, emphasizing its adaptability and potential for broader application.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Baurzhan Yessingeldinov

Limited Liability Partnership, Ascent Research Group

40 Bukhar Zhyrau Street, Astana City 010000, Kazakhstan

Email: yessingeldinov.b@gmail.com

1. INTRODUCTION

Since 2016, Kazakhstan has been implementing an attestation (appraisal) model that involves teachers examining their practice by identifying professional development areas and ways to improve them. Teacher appraisal, as a process for assigning a qualification category, can significantly influence teachers' professional development [1]. Appraisal procedures include qualification assessment, assessment of teachers' knowledge (ATK), and portfolio assessment.

Despite the aim to ensure a comprehensive approach to the appraisal process, in practice, achieving the threshold level of assessment of teachers' knowledge in most cases determines the success of appraisal. Recent studies in Kazakhstan reveal that teachers and school leaders prioritize summative indicators of their professional activities, such as test results and evidence collection of teacher and student achievements, over

fostering genuine learning [2], [3]. Consequently, the appraisal process does not realize the potential for teachers' development, favoring test results. This, in turn, shapes teachers' perception of appraisal, distorting its focus on professional development. Principles of constructivist (formative) and neoliberal (based on a market interpretation of "certification-salary increase") models are indeed incorporated into the policy and practice of teacher appraisal [3]. However, the neoliberal model of teacher evaluation is the key emphasis. Teachers prioritize the financial aspect in this context when determining the appraisal focus. The paramount concern about the efficacy of appraisal processes lies in aligning the categorization of qualifications as determined through appraisal outcomes and the caliber of educational practices. This alignment engenders heightened interest within societal domains and emerges as a focal point of scholarly discourse.

The identified contradictions define the research problem: despite the stated goal, teacher appraisal inadequately realizes its formative focus—teachers' professional development. The problem is exacerbated by insufficient research; in Kazakhstan, only isolated studies on small samples have been conducted in recent years, making it impossible to assess the effectiveness of appraisal in a broader context at school practice and educational policy levels. Therefore, developing a scientifically justified methodology to assess the impact of appraisal on teachers' professional development becomes relevant in verifying the validity and reliability of the evaluation procedures and tools used. In this sense, considering universally recognized definitions [4], validity in the research is understood as the degree to which evidence and theory support interpretations of appraisal results to determine the level of teachers' qualifications and reliability—the degree to which appraisal results are accurate and consistent across diverse evaluation instances. To address the research problem and purpose, the study states the following research question: what is the theoretically grounded content of a methodology aimed at assessing the impact of appraisal on teachers' professional development?

2. THE COMPREHENSIVE THEORETICAL BASIS

The effectiveness of educational reforms in Kazakhstan is closely tied to teacher proficiency, making professional development a key policy focus. Recent studies highlight the need for a deeper analysis of teaching performance to unlock potential and shape development trajectories. Comprehensive evaluation of teacher activities within the framework of teacher appraisal is essential. However, as noted by Ablayeva [3], the connection between appraisal and professional development is underrepresented in scientific discourse, causing appraisal to be seen as ineffective by teachers and administrators. Internationally, approaches to teacher appraisal vary in terms of goals, processes, and outcomes.

Teacher appraisal is often understood as assessing the effectiveness of teachers' activities to make judgments and provide feedback on their competencies and performance [5]. Appraisal can become a tool for ensuring quality (compliance with standards), stimulating teachers' reflection on their own practices, and providing information to support schools, teachers, and educational authorities in implementing educational policies [6]. However, the literature also emphasizes the contradictions between teacher appraisal's formative purposes, which are aimed at professional development [7], and summative purposes, which are related to accountability and managerial decisions [8]. Some authors argue that these are incompatible purposes, while others advocate for integrating them into the same teacher evaluation system.

A review of high-performing countries (Australia, Canada, Finland, Shanghai, and Singapore) shows continuous assessment and feedback are crucial for improving teacher performance [9]. In Singapore, the teacher career structure highlights the importance of appraisal in enhancing qualifications. Fairness and clarity in criteria are key to teacher satisfaction with appraisal processes and workplace motivation [10]. When appraisal processes are valid and reliable, and appraisers are competent, teacher experiences become more constructive, supported by school culture promoting professional development [6], [11]. Thus, validity and reliability are critical factors in ensuring the quality of the appraisal process.

Validity is the degree to which evidence and theory support interpreting measurement results for their intended use [12]. Based on Messick's validity theory, construct validity is foundational, emphasizing the need to validate all interpretations of test results [13]–[15]. Validation provides a scientific basis through accumulated evidence [4]. Reliability is the accuracy and consistency of assessments across different occasions [11], reflecting the consistency of repeated results. Classical test theory defines reliability through the correlation between scores on equivalent test forms. The importance of reliability grows with the significance of decision-making [4].

One widely used concept in the educational measurement community is Kane's argument-based approach to validity [16], [17]. Kane considers validity as the result of argumentation, which is built on the accumulation of theoretical foundations and corresponding evidence, unlike the traditional understanding of validity as a static assessment characteristic [18]. Therefore, Kane's approach [18] shifts the focus from a single validity characteristic to a more holistic view of validity (construct validity) and proposes a universal structure consisting of five claims for understanding and establishing validity.

Assessment of observed performance-assumes obtaining a formalized result (e.g., score). Generalization of observed results to the assessment domain (test domain)-implies extending the interpretation from evaluating results from a limited sample of observations (assessment of a specific set of indicators) to the expected value in the test domain (claims about the expected candidate efficiency in the assessment domain) [19], [20]. In other words, the score does not change, but its interpretation is expanded. Extrapolation from the assessment domain (test domain) to the domain of knowledge, skills, and judgment (KSJ)-assumes a prescription used to extend the interpretation from expected performance results on test tasks to expected results in the context of the domain of KSJ. This is an important aspect because certification exams use standardized, decontextualized versions of tasks in the domain of KSJ, which typically differ from tasks in the context of real professional activity. Therefore, decisions based on exam performance are extrapolated to actions in the domain of KSJ. In this case, the score (test result) does not change, but the interpretation of the scores is extended to the domain of KSJ.

Extrapolation from the domain of competencies (KSJ) to the domain of practice involves prescribing the use of interpretation from the domain of competencies to professional activity. The central assumption is that competencies in the domain of KSJ are necessary for effective professional practice. Therefore, the inability to acquire these competencies would be a severe obstacle in practice and, consequently, limit the effectiveness of the specialist's work. This assumption expands interpretation without changing the assessment.

A decision on certification involves the authority to use specific assessments to determine a candidate's readiness for effective practical activity. This decision is based on assumptions about the possible outcomes (both anticipated and unanticipated) of these decisions and their associated values [13], [14], [17], [21]. Kane emphasizes several advantages of using argument-based interpretation in certification exams [16], [17]. First, qualitative arguments take precedence over quantitative ones, relying on expert judgment rather than statistical support. Second, the approach is pragmatic, focusing on verifying candidates' readiness for practical work by developing tools that assess practical competencies. Third, critical assessment often takes a negative tone, where low scores indicate deficiencies affecting professional performance, while high scores do not necessarily reflect exceptional skills [22]. Additionally, three key rules guide argument-based validation: precise interpretation formulation, viewing validation as a comprehensive research program, and critically evaluating interpretations and possible refutations [23]. Interpretive arguments are dynamic and subject to revision, not solely determined by assessment tools or assumptions. Clear argumentation is crucial, as similar goals can stem from diverse claims. Kane warns against overly ambitious interpretations of test results, advocating for a more cautious approach in assessing professional readiness [24].

Effective validation is impossible without a clearly formulated interpretation and utilization of results since they are the ones being validated. To this end, Kane utilized Toulmin's model of inference for systematic and logical consideration of claims and evidence. The Toulmin model divides an argument into six parts [25]:

- Claim: interpretation/use of assessment results requiring justification.
- Data: evidence supporting the claim, such as measurement results and statistical analysis.
- Warrant: justification of why the data support the claim, such as theoretical models and research findings.
- Exception: a potential threat to validity, the condition under which the claim may be false or when the data or warrant may not be applicable.
- Backing: additional evidence or rationalization supporting the warrant.
- Qualifier: limitation on the strength (degree of certainty) of the claim, such as test reliability limits, demographic or contextual constraints, and other factors.

Thus, the literature review confirms that Kane's methodology provides theoretical justification for assessing the validity and reliability of the teacher appraisal process through interpretation and utilization of its results, emphasizing the importance of integrating a wide range of evidence. This methodology increases accuracy and objectivity in assessing teachers' professional qualifications. It expands the conceptual boundaries of teacher appraisal as a mechanism for professional development from a more constructive perspective. The Toulmin model provides a system for analyzing arguments, enabling Kane's claims about assessment validity to be justified. Both approaches emphasize the importance of clarity, consistency, and evidence in argumentation.

3. METHOD

The research question in this article is: what is the theoretically grounded content of a methodology aimed at assessing the impact of appraisal on teachers' professional development? The premise is that higher reliability and validity of appraisal processes lead to a more significant influence on development. Appraisal with high validity and reliability can stimulate teachers' growth by accurately assessing their alignment with established requirements. The hierarchical structure of appraisal helps define development goals and professional growth trajectories. To define the theoretical framework and address the research question, the following methods were employed:

The methodology for assessing the impact of appraisal on teachers' professional ... (Aidana Shilibekova)

- A comparative analysis of validation approaches considered validity as a test property and part of a holistic assessment design. The analysis of the argument-based [16], [17], evidence-centered [26], and validity framework [26] approaches identified Kane's argument-based approach [16], [17] as the most suitable for the current methodology. Kane's approach, which applies to high-stakes assessments, is adaptable to the complex appraisal process involving multiple procedures. Validity results from argumentation based on theoretical foundations and evidence rather than a static characteristic. This comprehensive view uses a universal structure of claims with three key rules: precise interpretation, validation as a research program, and critical evaluation of alternative explanations [18], [27].
- Reviewing appraisal materials in the context of qualitative research methods strengthens the scientific rigor of the conclusions. This review serves as an additional source of information and allows for data triangulation [28]. An adapted content analysis instrument with guiding questions was used in document analysis, where professional standards, appraisal rules, and teacher qualification requirements were considered to define the documents' policy.
- The methodology was modelled from the perspective that effective validation is impossible without a formulated interpretation and use of results, as these are the subject of validation. For this purpose, Kane [16], [17] utilized Toulmin's argumentation analysis model (Toulmin's model of inference) to examine claims and evidence systematically and logically. This allowed for determining the methods and tools for assessing the validity and reliability of each claim.

Overall, the methods employed confirmed that Kane's approach provides theoretical justification for assessing the validity and reliability of the teacher appraisal process through interpretation and use of results, emphasizing the significance of integrating a wide range of evidence. This approach increases the accuracy and objectivity of assessing teachers' professional qualifications. It expands the conceptual boundaries of appraisal as a mechanism for professional development from a more constructive perspective. The subsequent justification and description of the methodology's content focus on the analysis of methodological approaches, leaving the methods of its experimental verification and results analysis beyond the scope of this article.

4. RESULTS AND DISCUSSION

In Kazakhstan, teacher appraisal evaluates professional competencies based on assigned qualification categories. The process includes qualification assessment, ATK and essay writing, and a comprehensive analytical summary of teacher activities (portfolio assessment), all of which must be completed at least once every five years. A qualification category is granted only if the teacher receives positive assessments in all three areas. Due to its complexity, teacher appraisal can lead to diverse interpretations of validity, further complicated by the lack of research in this field. This highlights the need for a unique methodology to assess appraisal validity, incorporating international experience [29].

Kane's approach to developing the methodology offers several advantages. First, it prioritizes qualitative characteristics over quantitative ones through interpretational arguments, which is particularly useful given the complexity and variability of pedagogical activities. Second, the diversity of characteristics and subjects in appraisal procedures requires a differentiated approach to checking validity and reliability. By integrating the results of three procedures, the appraisal leads to a unified decision on qualification categories, emphasizing the complexity of the construct. Third, Kane's standardized validation technology supports the transition from individual methods to a theoretically grounded methodology, using sequential interpretational arguments aligned with the appraisal procedures.

4.1. The methodology for assessing the impact of appraisal on teachers' professional development

The methodology is developed using Kane's validity approach [16], [17] to determine methods for assessing the validity and reliability of appraisal procedures, corresponding instruments, and potential evidence. It also addresses the corresponding instrument and potential evidence. The methodology includes: i) formulation of claims (interpretive arguments) adapted to teacher appraisal as shown in Table 1; ii) description of adapted claims and methods for their assessment with potential supporting evidence; and iii) structural analysis of argumentation based on the Toulmin model applied to adapted claims as shown in Table 2.

It should be noted that differences in characteristics and assessment domains of the three appraisal procedures necessitated the application of a differentiated approach in the methodology, specifically: i) qualification assessment will not be considered in claims 3 and 4 due to its limitations (only the right to undergo appraisal) in extrapolating its decision to the KSJ and practice domains [16], [17]; ii) ATK will be considered in claim 3 for extrapolation from the test domain to the KSJ domain; and iii) portfolio assessment will be considered in claim 4 for extrapolation from the KSJ domain to the practice domain. The five adapted claims, evaluation methods, and potential supporting evidence are then examined in relation to teacher appraisal.

Table 1. Claims (interpretive arguments)

No.	Kane's claim	Claim adapted for teacher appraisal
1.	Evaluation of observed performance	According to established rules, appraisal procedures allow for the determination of the level of teachers' qualifications (evaluation) when assigning a qualification category.
2.	Generalization of the observed results to the assessment domain	Each of the three appraisal procedures allows for generalizing observed results to corresponding test domains: qualification requirements, subject and methodological knowledge (ATK domain), and teacher competencies (portfolio assessment domain).
3.	Extrapolation from the test domain to the KSJ domain	The ATK domain (test domain) allows for reflecting the domain of teacher competence (KSJ domain).
4.	Extrapolation from the KSJ domain to the practice domain.	The portfolio assessment domain (KSJ domain) allows for reflecting the domain of pedagogical activity (practice domain).
5.	Decision about readiness for practice	The decision to assign a qualification category—based on qualification assessment criteria, passing scores of ATK, and portfolio assessment—is a reliable and valid assessment of teachers' qualification levels, thereby contributing to their professional development.

4.1.1. Claim 1

According to established rules, appraisal procedures allow for determining the level of teachers' qualifications (evaluation) when assigning a qualification category. The final appraisal decision is based on the results – evaluation of observed performance - obtained from each procedure: qualification assessment, ATK, and portfolio assessment. Qualification assessment is conducted by reviewing documents submitted by the teacher to confirm compliance with formal qualification requirements: documents on education, retraining (if any), and work experience. Data collection is carried out from the information systems of relevant government bodies. Depending on the teacher's declared qualification category—the school or education management bodies—the authorized body reviews the application and issues a notification of acceptance or refusal to accept documents. The decision to accept or reject documents is considered an evaluation of observed performance in the qualification assessment procedure, i.e., the teacher's compliance with formal requirements.

An ATK is conducted through testing, consisting of two sections with multiple-choice questions: "subject content knowledge"-30 questions, and "teaching methodology"-20 questions. The test result is considered positive when threshold scores (% completion) are achieved according to the qualification category: "teacher" (50%), "teacher-moderator" (60%), "teacher-expert" (70%), "teacher-researcher" (80%), "teacher-master" (90%). Testing is conducted in a computerized format, with one point awarded for each correct answer. After reviewing the results, teachers can submit an appeal to the computerized testing system. During the appeal process, a group of subject matter experts reviews the teacher's application and decides. The correct answer is justified by information (facts) in textbooks and primary sources.

At the end of the test, the teacher shall write an essay (250-300 words) on a topic determined by the authorized body. At the same time, the essay is not evaluated and does not affect the decision-making on teacher appraisal; it cannot be considered as an evaluation of the observed performance. Thus, the observed score on the ATK is formed only based on test results.

A comprehensive analytical synthesis of teacher activity results (portfolio assessment) is conducted to determine if a teacher's practice meets qualification requirements. The portfolio is evaluated based on teaching quality, student achievements in competitions, teacher achievements in professional events, and implementation of best pedagogical practices. An expert council reviews portfolios according to the teacher's qualification category, assessing each in the teacher's presence. However, evidence of compliance with this norm was not found for all teachers. After the review, council members complete assessment forms, providing recommendations on the teacher's qualification status. If the portfolio does not meet the requirements, the teacher cannot proceed to the next stage (ATK).

The strength of the evaluative inference, or the degree of confidence that the obtained score indicates the candidate's answer quality, depends on how appropriate the scoring procedures were and how carefully and consistently they were applied [16], [17]. Accordingly, to determine the strength of evaluative inferences regarding the three appraisal procedures, it was assumed that: i) qualification assessment does not measure observed performance but provides a normative basis and prerequisites for teacher appraisal, establishing minimum tenure requirements for qualification categories. The methodology may assess the relevance of tenure as a criterion for qualification categories, its correlation with ATK threshold scores, and portfolio evidence; ii) the ATK offers an objective assessment through a test format with one correct answer. Scores are based on the assumption that no external factors affect their interpretation. If this assumption is challenged, adjustments to assessment procedures or interpretive arguments may be needed. The methodology includes document analysis to evaluate corrective norms and their impact on results; and iii) unlike test assessments, portfolio assessment is based on judgments, requiring additional reliability confirmation. To assess the consistency of decisions within the methodology, data collection and analysis are envisaged regarding the generalizing ability of assessments given by different experts and inter-territorial reliability.

In general, exceptions to claim 1 may arise from violations of established procedures. Assessment may be questioned if serious violations occurred during procedures (e.g., document falsification, technical glitches during testing, or expert negligence or bias). The discovery of evidence confirming such violations affects the strength of evaluative inferences and is reflected in the degree of certainty (qualifier).

4.1.2. Claim 2

Each of the three appraisal procedures allows for generalizing observed results to corresponding assessment domains (test domains)—qualification requirements, subject and methodological knowledge (ATK domain), and teacher competencies (portfolio assessment domain). The qualification assessment determines a teacher's compliance with appraisal requirements based on document submission, resulting in a clear yes/no decision. This assessment is repeated at each appraisal, considering any changes. The decision assumes that longer tenure correlates with a higher qualification category, though qualitative confirmation remains essential. The ATK results aim to generalize assessment across subject and methodological knowledge. Generalization is based on specific test tasks, often analyzed using reliability coefficients (alpha or G coefficient). This generalization is valid if the quality of assessment tools, procedures, and results processing is ensured.

Developing test structures and specifications and conducting pilot testing using statistical and psychometric results processing can ensure the quality of assessment tools. Standardizing the test format, structure, and timeframes reduces the variability of results associated with these parameters, thereby narrowing the test domain [16], [17]. Standardization, strict adherence to testing requirements, and reliable data collection ensure assessment quality. Invariance is generally assumed for diverse testing conditions, such as audience type or desk setup. These factors are expected not to impact performance within acceptable limits significantly.

The quality of results processing can be ensured by standardization using statistical and psychometric results processing, including calculating test reliability indicators. Using statistical adjustments to equate scores and control specific sources of errors is particularly justified when standardization is not possible, for example, when tasks cannot be reused in testing with high stakes. Justifying such equating procedures requires supporting the adequacy of the equating model, model-data fit, and evidence that equating errors are not too significant.

Thus, provided that qualitative assessment tools, procedures, and results processing are used, it is possible to verify how ATK allows generalizing results to assess the test domain through statistical indicators. The decision is based on portfolio assessment results and assessing evidence according to established criteria. Based on individual evidence of teacher practice, it is assumed that this decision (observed score) generalizes the assessment domain—pedagogical activities. Thus, the interpretation of portfolio assessment from limited observation evidence expands to the expected value in the assessment domain.

Research on the generalizability (or reliability) of observed results to the assessment domain plays a vital role in determining the accuracy of estimates of the expected solution (i.e., standard errors of measurement) and, consequently, in determining the strength of claims based on these estimates. Standard errors and confidence intervals are direct indicators of confidence that decisions based on the results of each assessment procedure (observed scores) are reliable and expected (fair) in the domains of their assessment—qualification requirements, subject and methodological knowledge, and pedagogical practice. To determine the generalizability (or reliability) of observed results to the assessment domain across three appraisal procedures, the following is assumed: i) for qualification assessment, surveying teachers regarding the perception of the degree of its representativeness, reliability, and significance for determining the level of qualification category; ii) for ATK (testing), checking the reliability of test results using appropriate statistical indicators; and iii) for portfolio assessment, expert judgment methods (on a random sample) should be used, considering the dynamic nature of pedagogical activities, the variability of potential evidence and their interpretations, and the conduct of assessment by an expert council. Additionally, it is recommended that teachers' perceptions be assessed through a survey to determine how reliable, fair, and sufficient the portfolio assessment process is for generalizing practice.

Exceptions to claim 2 arise for two main reasons: observations may not represent the assessment domain, or invariance may fail in specific cases. Serious violations in data collection procedures, such as insufficient test time, equipment failures, or breaches of academic integrity, can make decisions unrepresentative of the assessment domain. Additionally, while most errors are minor, certain instances like integrity breaches and conflicts of interest can lead to significant deviations.

4.1.3. Claim 3

The ATK domain (test domain) allows for reflecting the teacher competencies (KSJ domain). ATK should ensure the validity and reliability of the result in reflecting the teacher's KSJ. This claim assumes that

certification exam tasks, in the case of ATK, should be designed to reflect KSJ in practice. However, for most certification programs, the types of activities in the KSJ domain are complex and diverse (in the case of teacher appraisal, this could be teaching methodology), which in turn imposes significant limitations on the degree of confidence in this claim.

The strength with which one can assert about the KSJ domain based on ATK tests will depend on the overall confidence in the relationship between ATK results and KSJ results. This will depend on how well the ATK test tasks correspond to KSJ in practical activities and on empirical or theoretical evidence for or against this relationship. To determine this, it is recommended that ATK specifications be analyzed to ensure compliance with educational programs and documents defining the KSJ domain.

Understanding how teachers respond to ATK tasks and perform related pedagogical activities is key to arguing the plausibility of extrapolation, which relies more on expert opinions than formal models. Experts, such as administrators, mentors, and teachers, can provide insights through surveys or focus groups. The assumption is often negative, suggesting that teachers struggling with subject knowledge may underperform in pedagogy. Most evidence supporting extrapolation is negative, confirming the claim when no bias factors are identified. If doubt remains, empirical testing of significant factors may be necessary.

One approach to deepening the understanding of what the test items measure is collecting "think-aloud" protocols from teachers during ATK administration. These data can be obtained during individual sessions when researchers record a teacher's self-description of how they approach each task. Such data would provide a direct insight into how well teachers' test performance reflects their performance on corresponding tasks in the KSJ domain [27].

Extrapolation issues often fall under "construct-irrelevant variance" or "construct underrepresentation" [16], [17]. Construct-irrelevant variance arises when ATK tasks or response formats differ from those in the KSJ domain. Tasks with a single correct answer may lead to "guessing" and construct underrepresentation, where test answers do not reflect practical skills. To address this, reliable mechanisms for task development and quality assessment are needed, ensuring full KSJ domain coverage. Statistical methods should assess test discrimination, and specifications should be analyzed to justify using a single test for all qualification levels.

One aspect to consider could be the analysis of reporting on the results of ATK, particularly regarding the depth and coverage of result interpretation in terms of test structure according to specifications. Exceptions to the extrapolation inference are typically associated with cases where results in the testing domain likely systematically differ from results in the KSJ domain. Additionally, teachers' concerns about anxiety when taking ATK essentially serve as grounds for such exceptions. In a more general sense, any limiting factors (for example, disability) that hinder test performance but do not significantly affect results in the KSJ domain may, but not necessarily, lead to exceptions from the extrapolation inference.

4.1.4. Claim 4

The portfolio assessment domain (KSJ domain) allows for reflecting the domain of pedagogical practice. The second extrapolation extends the interpretation from the KSJ domain to the practice domain, based on the assumption that the assessed competencies through portfolio evidence play an essential role in teachers' practices. Therefore, claim 4 will consider to what extent portfolio assessment as an appraisal procedure allows for extrapolating the teacher's competencies to the practice domain. In this case, the basis for the extrapolation conclusion may be the use of the professional standard "Pedagogue," which defines teachers' competencies, content, quality, and conditions of their practices.

Teacher competency criteria are structured by qualification category progression, reflecting pedagogical practice requirements. During appraisal, evidence in the portfolio is expected to demonstrate competency as per the professional standard. If the portfolio lacks sufficient evidence, the teacher does not meet the qualification category, with negative argumentation prevailing. However, limitations include the possibility that the assessed competencies in the KSJ domain are not critical for effective pedagogy or that the KSJ domain covers too narrow a range, failing to represent key aspects of teaching. This limitation may challenge the extrapolation of portfolio results.

According to Kane [16], [17], the strength of asserting teacher performance based on portfolio evidence depends on the connection between the portfolio and pedagogical practices according to the professional standard. To verify the validity and reliability of portfolio assessments, methods include analyzing regulatory documents, checking portfolio evidence against the standard, assessing expert decision consistency, and conducting teacher surveys and focus groups. Since negative argumentation (teachers lacking KSJ competencies will not perform well) is stronger than positive, negatively framed questions are recommended for surveys and focus groups. Exceptions to the extrapolation inference may be related to questions regarding the quality of portfolio evidence, which may not cover the entire scope of the teachers' practices or may not correspond to the claimed qualification category.

Table 2. Structural analysis of argumentation of adapted claims according to the Toulmin model

Claim	Confirmation/refutation of claims	
	Warrant	Exception
Claim 1: According to established rules, appraisal procedures allow for determining the level of teachers' qualifications (evaluation) when assigning a qualification category.	Qualification assessment establishes the compliance of the teachers' documents with the requirements. Backing: a procedure for automated data collection. Qualifier: a strong degree. ATK (and essay writing) establishes the subject and methodological knowledge level. Backing: computerized testing on subject and methodology (including an essay) and appeal process. Qualifier: a moderate degree. A portfolio assessment verifies teacher compliance with qualification requirements. Backing: collegial decision-making process. Qualifier: a moderate degree.	The qualification assessment may risk verifying document completeness and authenticity. Backing: lack of full-fledged authenticity check, differentiation by quality and level of documents. Qualifier: a weak degree. The ATK procedure (including essay writing) may risk assessment tool quality (validity, reliability, and objectivity). Backing: statistical analysis in tool development and raw score calculation (no essay evaluation). Qualifier: a strong degree. The portfolio assessment may risk objectivity. Backing: lacks standardization, expert generalizability, and inter-territorial reliability. Qualifier: a strong degree.
Claim 2: Each of the three appraisal procedures allows for generalizing observed results to corresponding assessment domains (test domains)–qualification requirements, subject and methodological knowledge (ATK domain), and teacher competencies (portfolio assessment domain).	Qualification assessment summarizes checks of teacher documents for qualification requirements. Backing: established list of documents. Qualifier: a moderate degree. ATK generalizes testing results in subject and methodological knowledge. Backing: ATK specifications, reliability indicators, and task approbation. Qualifier: a moderate degree. Portfolio assessment summarizes expert evaluations of teacher competencies. Backing: approved criteria covering teaching quality, student achievements, and experience dissemination. Qualifier: a strong degree.	Qualification assessment provides only a normative basis for appraisal. Backing: formal document requirements. Qualifier: a strong degree. As a standardized test, ATK may limit the scope of subject and methodological knowledge. Backing: no approach for covering training programs in ATK specifications and weak reliability indicators. Qualifier: a strong degree. Portfolio assessment may reduce decision reliability due to complex teacher competencies, causing variability in evidence interpretation. Backing: rubric weakly differentiating qualification categories and low expert agreement reliability. Qualifier: a moderate degree.
Claim 3: The ATK domain (test domain) allows for reflecting the teacher competencies (KSJ domain).	The ATK allows reflection on a teacher's domain of subject and methodological competencies. Backing: The "Subject Content Knowledge" test specification is based on the subject curriculum, while the "Teaching Methodology" test is aligned with teaching materials per the curriculum and ATK task approval. Qualifier: a moderate degree.	The ATK may not align with standards defining teacher competence. Backing: ATK specifications lack reference to teacher qualification characteristics and professional standards. Qualifier: a strong degree. ATK content may lack differentiation by teacher qualification levels (same test for all). Backing: Passing thresholds by qualification categories. Qualifier: a moderate degree.
Claim 4: The portfolio assessment domain (KSJ domain) allows for reflecting the domain of pedagogical practice.	Evaluating portfolio evidence per professional standard requirements reflects pedagogical practice. Backing: appraisal rules and expert evaluations. Qualifier: a moderate degree. Portfolio assessment uses practice-based evidence to reflect the teaching domain. Backing: assessment criteria, lesson observation sheet. Qualifier: a moderate degree.	Complexity in teacher competencies and leveling may risk covering key practice areas in portfolio assessment. Backing: list of competencies in the professional standard. Qualifier: a moderate degree. Lack of standardized portfolio interpretation may bias assessments of teacher competencies. Backing: Expert council minutes, evaluation sheet. Qualifier: a strong degree.
Claim 5: The decision to assign a qualification category based on qualification assessment criteria, passing scores of ATK tests, and portfolio assessment is considered a reliable and valid assessment of teachers' qualification levels, thereby contributing to their professional development.	The three appraisal outcomes collectively determine teachers' compliance with qualification standards. Backing: teacher qualification criteria, appraisal rules, ATK specifications, and expert evaluations. Qualifier: a moderate degree. Final evaluation decisions are shaped by constructive appraisals, fostering teacher growth and development. Backing: portfolios, including feedback in lesson observation and evaluation sheets. Qualifier: a moderate degree.	Dominance of one appraisal procedure may unbalance qualification decisions. Backing: statistical indicators on passing the appraisal in the context of appraisal procedures. Qualifier: a strong degree. Teachers may fail to confirm or downgrade qualification in later appraisals, raising reliability concerns. Backing: data on teachers who did not confirm or downgraded categories, with reasons analyzed. Qualifier: average degree. Assigning qualification categories may not guarantee improvements in teacher practice or student performance. Backing: weak correlation between teacher quality and student (UNT) results. Qualifier: a moderate degree.

4.1.5. Claim 5

The decision to assign a qualification category based on qualification assessment criteria, passing scores of ATK tests, and portfolio assessment is considered a reliable and valid assessment of teachers' qualification levels, thereby contributing to their professional development. It is assumed that when justifying decision-making rules, it is necessary to prove that the certification exam will achieve a specific goal (for example, societal protection) at reasonable costs. The existence of a specific goal provides grounds to believe that certification exams will, in any case, have a predominantly positive impact, even considering the costs [27]. In this sense, appraisal aims to thoroughly examine teachers' qualification levels to make informed decisions, ensuring professional development, career advancement, and appropriate remuneration for teachers and providing society with qualified personnel capable of delivering quality educational services. A thorough examination of teachers' qualification levels forms the basis for multi-procedural teacher appraisal: qualification assessment, ATK test, and portfolio assessment. As noted, positive decisions on each procedure are mandatory for the final decision on assigning the claimed qualification category.

The selection of criteria for qualification and portfolio assessments, along with the ATK passing score, is crucial in the certification process and must be justified. Thresholds should be high enough for societal protection but not overly restrictive [16], [17]. Empirical research, including surveys and focus groups with certification participants, can ensure the fairness and objectivity of these criteria and passing scores concerning qualification categories [30]. Limitations in decision-making may include assumptions of bias towards any individual or group. This may concern the legitimacy of one or both extrapolations or mainly focus on the choice of the passing score. Regarding appraisal, possibilities of bias may be subject to separate analysis through document review and organizational-technical conditions of procedure implementation. Exceptions to the decision-making process may arise for several reasons. For example, no factual data may confirm the decision, or teachers may be rejected if serious violations are discovered (fraud, plagiarism, and illegal conduct). As mentioned earlier, interpretive arguments may be unstable, and the decision-making rule may be refuted by additional evidence affecting the decision but not included in the rule. Further, in the methodology for conducting structural analysis based on the Toulmin model concerning adapted claims, the following primary data were used:

- i) For teacher appraisal: the professional standard, qualification characteristics of teacher positions, teacher appraisal rules, expert assessments of documents, commission meeting protocols for assigning qualification categories.
- ii) For qualification assessment: business process analysis, document lists, documentation for the information system.
- iii) For ATK testing: test results, reports on ATK test outcomes, specifications, sample questions, essay topics, procedural violation reports.
- iv) For portfolio assessment: assessment results, expert council protocols, assessment criteria, portfolio samples, portfolio assessment sheets, and lesson observation sheets.

Based on the analysis of this data, a structural analysis of the argumentation of adapted statements was formed according to S. Toulmin's model (Table 2). Grounds and counterarguments were formulated for the statements, and for each ground and counterargument, supports were determined to show the basis on which these grounds and counterarguments rest. Additionally, a qualifier indicating the level of certainty—graded as strong, medium, or weak—was established for each.

Using Kane's approach [16], [17] in assessing the validity of teacher appraisal allowed us to formulate interpretive arguments to be tested. We used a differentiated approach, where interpretive arguments were determined by separate procedures depending on the subject of evaluation. To verify the validity of each conclusion, the structure of argumentation analysis according to the Toulmin model was used, and the data collection methods were considered, which allowed the consistent and logical verification of the validity and reliability of conclusions and evidence.

The traditional view of teacher appraisal, focused on assigning qualification categories, is limited as it overlooks professional development needs and lacks constructive depth. Considering validity [16], [17] reshapes appraisal models, emphasizing its impact on teachers' development and evaluation quality. The proposed methodology allows for analyzing performance appraisal through aspects like observed performance, generalization, extrapolation, and validation, previously unused in Kazakhstan. Construct validity integrates other forms of validity, following Messick's multifaceted approach [13], [14], focusing on evidence supporting or refuting the interpretation of assessment results.

4.2. Implications and recommendations for further use

The methodology is aimed at assessing the validity and reliability of teacher appraisal procedures to determine the objectivity of the results of assigning qualification categories; therefore, it allows educational authorities to effectively analyze procedure compliance, identify critical points, and justify the directions of educational policy in the field of teacher professional development. The methodology can be adapted and

applied at different levels: school, district, region, and country. The structured algorithm that defines actions and indicates where to start and how to continue, the templates for tools, document analysis, and result interpretation make the methodology flexible and adaptable to different goals and educational contexts.

The methodology allows studying appraisal procedures using cross-sectional and longitudinal approaches. Cross-sectional studies focus on analyzing and assessing the condition at a particular point in time among different groups of participants. In contrast, longitudinal studies track progress and change in one group over a long period of time, providing insight not only into the immediate results of change but also its long-term impact on the professional development of teachers and the quality of the educational process.

The methodology emphasizes and expands teachers' professional development potential through appraisal, emphasizing formative goals and results. It forms a strategic and tactical vision necessary for thinking about the directions of development of teacher appraisal in general and procedures in particular (e.g., optimization and standardization of processes, creation of a feedback system). Such a vision at the level of teachers contributes to actualizing the processes of identifying their needs and launching mechanisms for managing their professional development.

The methodology assumes the involvement of specialists with sufficient competence in certifying and evaluating teachers' professional activity. In addition, it provides a methodological basis for developing educational programs aimed at training specialists and creating courses for teachers based on the findings obtained through this methodology. In the long term, this will enable forming an expert community to improve teacher appraisal and professional development.

5. CONCLUSION

This study provided a comprehensive evaluation of the impact of teacher appraisal procedures on their professional development in Kazakhstan. To do this, the paper addresses the following central question: What could be the theoretically grounded content of a methodology aimed at assessing the impact of appraisal on the professional development of teachers? The analysis identified key issues with the current appraisal system, particularly the overemphasis on test results, which hinders the full potential of these procedures to support teachers' professional growth. By applying Kane's approach to validity and Toulmin's model of argumentation, the validity and reliability of existing appraisal procedures were assessed. The findings demonstrated that improving these aspects could significantly enhance the effectiveness of the appraisal procedures, making them more objective and fairer. Key factors influencing the success of these procedures were identified, leading to specific recommendations for their improvement. This includes shifting the focus from testing to a more holistic and multifaceted evaluation of teachers' professional activities.

The significance of these findings extends beyond a single educational context. In Kazakhstan, the conclusions drawn from this study could contribute to developing more effective and transparent appraisal procedures that better support teacher development and, consequently, improve the overall quality of education. In a broader context, the proposed methodology could be adapted and applied in other countries facing similar challenges in their teacher appraisal systems, highlighting its versatility and potential for future application.

The study's findings underscore the importance of a critical approach to designing and implementing teacher appraisal procedures. For appraisals to truly support professional development, it is essential to consider quantitative metrics and the qualitative aspects of teachers' professional activities. In this context, the proposed methodology provides tools for a deeper and more comprehensive analysis of appraisal processes, opening up new possibilities for improvement.

Future research in this area could lead to even more precise and practical assessment methods that contribute to creating fair and efficient appraisal systems. This, in turn, would positively impact the education system as a whole, ensuring higher quality teaching and learning. Therefore, the conclusions and recommendations of this study can be seen as an important step towards the development of a more equitable and effective teacher appraisal system, both in Kazakhstan and beyond.

ACKNOWLEDGEMENTS




This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19679296).

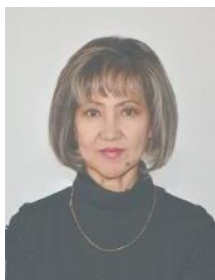
REFERENCES




- [1] A. Zhumykbayeva and M. Ablayeva, "Teacher attestation: identifying the factors influencing teacher reflective skills (in Russian)," *BULLETIN Series of Pedagogical Sciences*, vol. 79, no. 3(2023), pp. 256–264, Nov. 2023, doi: 10.51889/2959-5762.2023.79.3.022.
- [2] N. Ayubayeva, "Teacher collaboration for professional learning: case studies of three schools in Kazakhstan," Ph.D. dissertation, University of Cambridge, United Kingdom, 2018, doi: 10.17863/CAM.20729.
- [3] M. Ablayeva, "Teacher appraisal for professional learning: perspectives from one secondary school in Kazakhstan," Ph.D. dissertation, Nazarbayev University, Republic of Kazakhstan, 2022.
- [4] AERA, APA, and NCME, *Standards for educational and psychological testing: national council on measurement in education*. Washington DC: American Educational Research Association, 2014.
- [5] T. T. H. Nguyen and M. Walker, "Sustainable assessment for lifelong learning," *Assessment & Evaluation in Higher Education*, vol. 41, no. 1, pp. 97–111, Jan. 2016, doi: 10.1080/02602938.2014.985632.
- [6] Organization for Economic Co-operation and Development (OECD), *Education policy outlook: Kazakhstan*. OECD Publishing, 2018.
- [7] L. Darling-Hammond, *Powerful teacher education: lessons from exemplary programs*. San Francisco, CA: John Wiley & Sons, Inc., 2012.
- [8] M. A. Flores, "Teacher evaluation in Portugal: persisting challenges and perceived effects," *Teachers and Teaching*, vol. 24, no. 3, pp. 223–245, Apr. 2018, doi: 10.1080/13540602.2018.1425677.
- [9] L. Darling-Hammond, M. E. Hyler, and M. Gardner, *Effective teacher professional development*. Palo Alto, CA: Learning Policy Institute, 2017.
- [10] K. O. Kelly, S. Y. A. Ang, W. L. Chong, and W. S. Hu, "Teacher appraisal and its outcomes in Singapore primary schools," *Journal of Educational Administration*, vol. 46, no. 1, pp. 39–54, Feb. 2008, doi: 10.1108/09578230810849808.
- [11] M. Tuytens, N. Moolenaar, A. Daly, and G. Devos, "Teachers' informal feedback seeking towards the school leadership team. A social network analysis in secondary schools," *Research Papers in Education*, vol. 34, no. 4, pp. 405–424, Jul. 2019, doi: 10.1080/02671522.2018.1452961.
- [12] N. Bukhari, "Comparisons of and concerns about two testing application chapters in the 2014 standards for educational and psychological testing," *Universal Journal of Educational Research*, vol. 8, no. 10, pp. 4603–4609, Oct. 2020, doi: 10.13189/ujer.2020.081028.
- [13] S. Messick, "Validity," in *Educational Measurement*, 3rd ed., R. Linn, Ed. New York, NY: American Council on Education and Macmillan Publishing Company, 1989, pp. 13–103.
- [14] S. Messick, "Trait equivalence as construct validity of score interpretation across multiple methods of measurement," in *Construction Versus Choice in Cognitive Measurement*, New York, NY: Routledge, 2012, pp. 73–86, doi: 10.4324/9780203052518-7.
- [15] S. Lane, "Test-based accountability systems: the importance of paying attention to consequences," *ETS Research Report Series*, vol. 2020, no. 1, pp. 1–22, Dec. 2020, doi: 10.1002/ets2.12283.
- [16] M. Kane, "Certification testing as an illustration of argument-based validation," *Measurement: Interdisciplinary Research & Perspective*, vol. 2, no. 3, pp. 135–170, Jul. 2004, doi: 10.1207/s15366359mea0203_1.
- [17] M. T. Kane, "Validating the interpretations and uses of test scores," *Journal of Educational Measurement*, vol. 50, no. 1, pp. 1–73, Mar. 2013, doi: 10.1111/jedm.12000.
- [18] T. Kumazawa, T. Shizuka, M. Mochizuki, and A. Mizumoto, "Validity argument for the VELC Test® score interpretations and uses," *Language Testing in Asia*, vol. 6, no. 1, p. 2, Dec. 2016, doi: 10.1186/s40468-015-0023-3.
- [19] A. Robitzsch and O. Lüdtke, "Linking errors in international large-scale assessments: calculation of standard errors for trend estimation," *Assessment in Education: Principles, Policy & Practice*, vol. 26, no. 4, pp. 444–465, Jul. 2019, doi: 10.1080/0969594X.2018.1433633.
- [20] W. P. Vispoel, C. A. Morris, and M. Kilinc, "Using generalizability theory with continuous latent response variables," *Psychological Methods*, vol. 24, no. 2, pp. 153–178, Apr. 2019, doi: 10.1037/met0000177.
- [21] C. Rapanta, M. Garcia-Mila, and S. Gilabert, "What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education," *Review of Educational Research*, vol. 83, no. 4, pp. 483–520, Dec. 2013, doi: 10.3102/0034654313487606.
- [22] E. Amaral and J. Norcini, "Quality assurance in health professions education: role of accreditation and licensure," *Medical Education*, vol. 57, no. 1, pp. 40–48, Jan. 2023, doi: 10.1111/medu.14880.
- [23] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests," in *Research Design*, 1st ed., New York, NY: Routledge, 2017, pp. 225–238, doi: 10.4324/9781315128498-18.
- [24] M. Neumann, A. S. M. Niessen, and R. R. Meijer, "Implementing evidence-based assessment and selection in organizations: a review and an agenda for future research," *Organizational Psychology Review*, vol. 11, no. 3, pp. 205–239, Aug. 2021, doi: 10.1177/2041386620983419.
- [25] S. Erduran, "Toulmin's argument pattern as a 'horizon of possibilities' in the study of argumentation in science education," *Cultural Studies of Science Education*, vol. 13, no. 4, pp. 1091–1099, Dec. 2018, doi: 10.1007/s11422-017-9847-8.
- [26] M. E. Oliveri, R. Lawless, and R. J. Mislevy, "Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments," *International Journal of Testing*, vol. 19, no. 3, pp. 270–300, Jul. 2019, doi: 10.1080/15305058.2018.1543308.
- [27] V. Bademci, "Correcting fallacies about validity as the most fundamental concept in educational and psychological measurement," *International e-Journal of Educational Studies*, vol. 6, no. 12, pp. 148–154, Nov. 2022, doi: 10.31458/iejes.1140672.
- [28] N. Jentoft and T. S. Olsen, "Against the flow in data collection: how data triangulation combined with a 'slow' interview technique enriches data," *Qualitative Social Work*, vol. 18, no. 2, pp. 179–193, Mar. 2019, doi: 10.1177/1473325017712581.
- [29] C. Herlihy et al., "State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems," *Teachers College Record: The Voice of Scholarship in Education*, vol. 116, no. 1, pp. 1–28, Jan. 2014, doi: 10.1177/016146811411600108.
- [30] R. Coudret, S. Girard, and J. Saracco, "A new sliced inverse regression method for multivariate response," *Computational Statistics & Data Analysis*, vol. 77, pp. 285–299, Sep. 2014, doi: 10.1016/j.csda.2014.03.006.

BIOGRAPHIES OF AUTHORS






Aidana Shilibekova    is a candidate of Pedagogical Sciences, National Center for Teacher Professional Development “Orleu”. Her area of research is education, assessment and teacher professional development. She can be contacted at email: shilibekova_a@outlook.com.






Saule Vildanova    is a Master of Humanitarian Sciences National, National Center for Teacher Professional Development “Orleu”. Her area of research is education, assessment and teacher professional development. She can be contacted at email: svildanova@orleu-edu.kz.






Venera Mussarova    is a Master of Biology, National Center for Teacher Professional Development “Orleu”. Her area of research is education, teaching biology, assessment and teacher professional development. She can be contacted at email: vmussarova@orleu-edu.kz.






Baurzhan Yessingeldinov    is a Ph.D., Ascent Research Group. His area of research is education, teaching mathematics, assessment and teacher professional development. He can be contacted at email: yessingeldinov.b@gmail.com.



Moldir Ablayeva    is a Ph.D., National Center for Teacher Professional Development “Orleu”. She holds a Ph.D. in Education and an M.Sc. in Educational Leadership from Nazarbayev University, Astana, Kazakhstan. Her doctoral dissertation focused on “Teacher appraisal for professional learning: perspectives from one secondary school in Kazakhstan.” She can be contacted at email: ablayeva_m@orleu-edu.kz.



Amina Kaldybek    is a fourth-year sociology student at Nazarbayev University, an Autonomous Educational Organization. Her area of research is education, social inequality and postcolonialism. She can be contacted at email: amina.kaldybek@nu.edu.kz.